Article

Univariate L^p and l^p Averaging, $0 < p < 1$, in Polynomial Time by Utilization of Statistical Structure

John E. Lavery [1,2]

[1] Mathematical Sciences Division and Computing Sciences Division, Army Research Office, Army Research Laboratory, P.O. Box 12211, Research Triangle Park, NC 27709-2211, USA; E-Mail: john.e.lavery4.civ@mail.mil; Tel.: +1-919-549-4253; Fax: +1-919-549-4354

[2] Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC 27695-7906, USA

Abstract: We present evidence that one can calculate generically combinatorially expensive L^p and l^p averages, $0 < p < 1$, in polynomial time by restricting the data to come from a wide class of statistical distributions. Our approach differs from the approaches in the previous literature, which are based on *a priori* sparsity requirements or on accepting a local minimum as a replacement for a global minimum. The functionals by which L^p averages are calculated are not convex but are radially monotonic and the functionals by which l^p averages are calculated are nearly so, which are the keys to solvability in polynomial time. Analytical results for symmetric, radially monotonic univariate distributions are presented. An algorithm for univariate l^p averaging is presented. Computational results for a Gaussian distribution, a class of symmetric heavy-tailed distributions and a class of asymmetric heavy-tailed distributions are presented. Many phenomena in human-based areas are increasingly known to be represented by data that have large numbers of outliers and belong to very heavy-tailed distributions. When tails of distributions are so heavy that even medians (L^1 and l^1 averages) do not exist, one needs to consider using l^p minimization principles with $0 < p < 1$.

Keywords: average; heavy-tailed distribution; L^p average; l^p average; median; mode; polynomial time; radial monotonicity; statistical structure; univariate

1. REPORT DATE **2012**	2. REPORT TYPE	3. DATES COVERED **00-00-2012 to 00-00-2012**

4. TITLE AND SUBTITLE **Univariate Lp and lp Averaging, 0 < p < 1, in Polynomial Time by Utilization of Statistical Structure**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Army Research Office,Army Research Laboratory,P.O. Box 1221,Research Triangle Park,NC,27709**	8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Algorithms 2012, 5, 421-432; Published: 5 October 2012

14. ABSTRACT
We present evidence that one can calculate generically combinatorially expensive Lp and lp averages, 0 < p < 1, in polynomial time by restricting the data to come from a wide class of statistical distributions. Our approach differs from the approaches in the previous literature, which are based on a priori sparsity requirements or on accepting a local minimum as a replacement for a global minimum. The functionals by which Lp averages are calculated are not convex but are radially monotonic and the functionals by which lp averages are calculated are nearly so, which are the keys to solvability in polynomial time. Analytical results for symmetric, radially monotonic univariate distributions are presented. An algorithm for univariate lp averaging is presented. Computational results for a Gaussian distribution, a class of symmetric heavy-tailed distributions and a class of asymmetric heavy-tailed distributions are presented. Many phenomena in human-based areas are increasingly known to be represented by data that have large numbers of outliers and belong to very heavy-tailed distributions. When tails of distributions are so heavy that even medians (L1 and l1 averages) do not exist, one needs to consider using lp minimization principles with 0 < p < 1.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**13**	

1. Introduction

Minimization principles based on the l^1 and L^1 norms have recently rapidly become more common due to discovery of their important roles in sparse representation in signal and image processing [1,2], compressive sensing [3,4], shape-preserving geometric modeling [5,6] and robust principal component analysis [7–9]. In compressive sensing and sparse representation, it is known that, under proper sparsity conditions (for example, the restricted isometry property [3,4]), l^1 solutions are equivalent to "l^0 solutions", that is, the sparsest solutions, an important result because it allows one to find the solution of a combinatorially expensive l^0 maximum-sparsity minimization problem by a polynomial-time linear programming procedure for minimizing l^1 functionals. When the data follow heavy-tailed statistical distributions and the tails of the distributions are "not too heavy," various l^1 minimization principles, in the form of calculation of medians and quantiles, are primary choices that are efficient and robust against the many outliers [10–12]. Such distributions correspond to the uncertainty in many human-based phenomena and activities, including the Internet [13,14], finance [15,16] and other human and physical phenomena [16]. l^1 minimization principles are applicable also to data from light-tailed distributions such as the Gaussian, but, for such distributions, are less efficient than classical procedures (calculation of standard averages and variances).

When tails of the distributions are so heavy that even l^1 minimization principles do not exist, one needs to consider using l^p minimization principles with $0 < p < 1$, a topic on which investigation has recently started [2,3,17–20]. l^p minimization principles, $0 < p < 1$, are of interest because they produce solutions that are in general sparser, that is, closer to l^0 solutions, than l^1 minimization principles [20]. However, when $0 < p < 1$, solving l^p minimization principles is generically combinatorially expensive (NP-hard) [18], because l^p minimization principles can have arbitrarily large numbers of local minima. ("Generically" means "in the absence of additional information.") Investigations about polynomial-time l^p minimization, $0 < p < 1$, have focused on (1) obtaining local rather than global solutions [2,18,20] and (2) achieving a global minimum by restricting the class of problems to those with sufficient sparsity [3,17,19] (the approach used in compressive sensing). However, local solutions often differ strongly from global solutions and sparsity restrictions are often not applicable. The fact that the l^0 solution is, relative to other potential solutions, the sparsest solution does not imply that this solution is sparse to any specific degree. The sparsest solution may not be sparse in any absolute sense at all; it is just sparser than any other solution.

The approach that we will investigate in the present paper shares with compressive sensing the strategy of restricting the nature of the problem to achieve polynomial-time performance. However, we do so not by requiring sparsity to some *a priori* set level but rather by restricting the data to come from a wide class of statistical distributions, an approach not previously considered in the literature. This restriction turns out to be mild, often verifiable and often realistic since the problem as posed is often meaningful only when the data come from a statistical distribution. The approach in this paper differs from the approaches in the previous literature on l^p minimization principles also in a second way, namely, in that it starts the investigation of l^p minimization principles from consideration of their continuum analogues, L^p minimization principles.

The classes of L^p and l^p minimization principles that we will investigate in this paper are those that represent univariate continuum L^p averaging and discrete l^p averaging, defined as follows. Univariate

L^p and l^p averages are the real numbers a at which the following functionals A and B achieve their respective global minima:

$$A(a) := \int_{-\infty}^{\infty} |x - a|^p \, \psi(x) \, dx \tag{1}$$

where ψ is a probability density function (pdf) that satisfies the conditions given below, and

$$B(a) := \sum_i |x_i - a|^p \tag{2}$$

where the x_i are data points from the distribution with pdf ψ. The pdf ψ is assumed to have measurable second derivative and to satisfy the following two conditions:

- radially strictly monotonically decreasing outwards from the mode (3a)
- ψ and $d\psi/dx$ bounded by $c|x|^{-\beta}$ and $c|x|^{-\beta-1}$, respectively, for given c and $\beta > p + 1$ as $x \to \pm\infty$ (3b)

Without loss of generality, we assume that the mode, that is, the x at which ψ achieves its maximum, is at the origin.

In a departure from the traditional use of x as the independent variable of a univariate pdf, we will express univariate pdfs in radial form with r being the radius measured outward from the mode of the distribution. (This notation is chosen to allow natural generalization to higher dimensions in the future.) With the notation $g(r) = \psi(-r)$ and $f(r) = \psi(r)$, $r \geq 0$, functional A can be rewritten in the form

$$A(a) = \int_0^{\infty} |r + a|^p \, g(r) \, dr + \int_0^{\infty} |r - a|^p \, f(r) \, dr \tag{4}$$

Since functional (4) is finite only when

$$p < \beta - 1 \tag{5}$$

the mean (L^2 average) does not exist for distributions with $\beta \leq 3$ and even the median (L^1 average) does not exist for distributions with $\beta \leq 2$. For example, the median does not exist for the Student t distribution with one degree of freedom because $\beta = 2$ for this distribution. To create meaningful "averages" in these cases, weighted and trimmed sample means have been proposed with success [21]. However, weighted and trimmed sample means require *a priori* knowledge of the specific distribution and/or of various parameters, knowledge that is often not available. Minimization of the L^p functional (4) or of the l^p functional (2) is, when $0 < p < \min\{1, \beta-1\}$, an alternative for creating an "average" for a heavy-tailed distribution or of a sample thereof.

In the present paper, we will investigate whether, by providing only the information that the data come from a "standard" statistical distribution that satisfies Conditions (3), the L^p and l^p averaging functionals A and B can be minimized in a way that leads to polynomial-time minimization of general L^p and l^p functionals. Specifically, in the next two sections, we will investigate to what extent the L^p and l^p averaging functionals are devoid of local minima other than the global minimum, a key feature in this process. For illustration of the theoretical results, we will present computational results for the following three types of distributions:

Distribution 1: Gaussian (light-tailed distribution) distribution with probability density function

$$f(r) = g(r) = \frac{1}{\sqrt{2\pi}} exp\left(\frac{-r^2}{2}\right) \tag{6}$$

Distribution 2: Symmetric heavy-tailed distribution with probability density function

$$f(r) = g(r) = \frac{1}{c}\left[\left(1 + \frac{\alpha}{2}\right) - \frac{\alpha}{2}r^2\right], 0 \leq r \leq 1 \tag{7a}$$

$$f(r) = g(r) = \frac{1}{c}r^{-\alpha}, 1 < r < \infty \tag{7b}$$

where

$$c = 2\left[\frac{\alpha}{\alpha - 1} + \frac{\alpha}{3}\right] \tag{7c}$$

(For Distribution 2, the β of condition (3b) is α.)

Distribution 3: Asymmetric heavy-tailed distribution with probability density function

$$g(r) = \frac{1}{c}\left[\left(1 + \frac{\alpha}{2}\right) - \frac{\alpha}{2}r^2\right], f(r) = \frac{1}{c}\left[\left(1 + \frac{\alpha}{2}\right) + \left(\frac{1}{2} - \alpha\right)r^2 + \left(\frac{\alpha - 1}{2}\right)r^3\right], 0 \leq r \leq 1 \tag{8a}$$

$$g(r) = \frac{1}{c}r^{-\alpha}, f(r) = \frac{1}{c}r^{-(1+\alpha)/2}, 1 < r < \infty \tag{8b}$$

(right tail heavier than left tail), where

$$c = \frac{49}{24} + \frac{5\alpha}{8} + \frac{3}{\alpha - 1} \tag{8c}$$

(For Distribution 3, the β of condition (3b) is $(1 + \alpha)/2$.)

In Distributions 2 and 3, α is a real number > 1. Gaussian Distribution 1 is used to show that the results discussed here are applicable not only to heavy-tailed distributions but also to light-tailed distributions. These results are applicable *a fortiori* to compact distributions with no tails at all (tails uniformly 0). (Analysis and computations were carried out with the uniform distribution and with a pyramidal distribution, two distributions with no tails, but these results will not be discussed here.) While L^p and l^p averages can be calculated for light-tailed and no-tailed distributions, there are more meaningful and more efficient ways, for example, arithmetic averaging, to calculate central points of light-tailed and no-tailed distributions. L^p and l^p averages are most meaningful for heavy-tailed distributions.
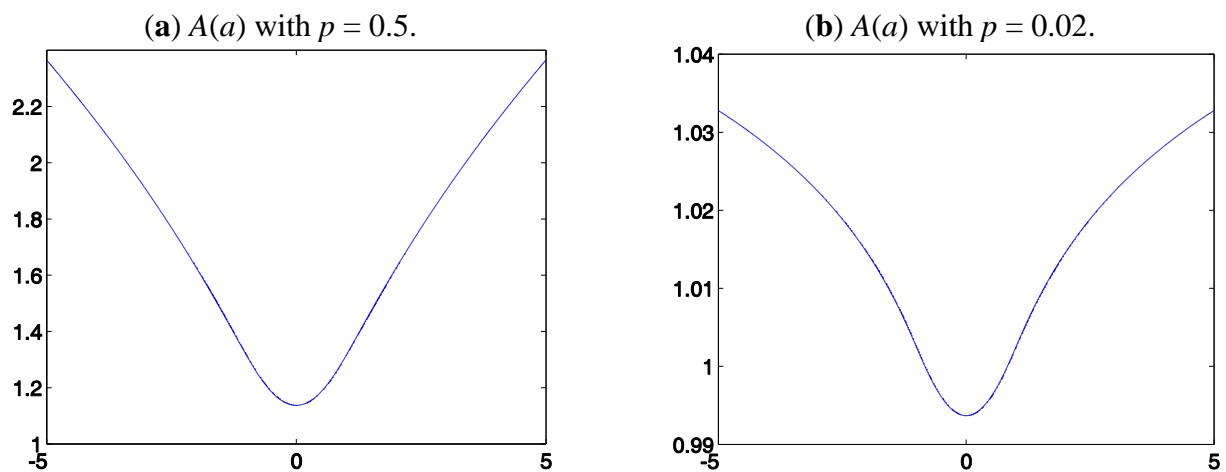
2. L^p Averaging

We present in Figures 1–3 the functionals $A(a)$ for Distributions 1–3, respectively, for various p. These functionals $A(a)$ have one global minimum at or near $r = 0$, no additional minima, are convex in a neighborhood of the global minimum and are concave outside of this neighborhood. The fact that the $A(a)$ are not globally convex is not important. Each $A(a)$ is radially monotonically increasing outward from its minimum, which is sufficient to guarantee that there is only one global minimum and that there are no other local minima. On every finite closed interval in Figures 1–3 that does not include the global minimum, the derivative dA/da is bounded away from 0. Hence, in all these cases, standard line-search methods converge to the global minimum in polynomial time. The structure of $A(a)$ seen in Figures 1–3 is due to the fact that $A(a)$ is based on a probability density function with strictly monotonically decreasing density in the radial directions outward from the mode. This structure does not generically occur for density functions $f(r)$ and $g(r)$ representing, for example, irregular scattered clusters. However, averaging in general and L^p averaging in particular make little sense when the data are clustered irregularly. The computational results presented in Figures 1–3 suggest the hypothesis

that, under "normal" statistical conditions on the data, L^p averaging is well posed and computationally tractable. In the remainder of this section, we will investigate portions of this hypothesis.

Figures 1. L^p averaging functional $A(a)$ for Gaussian Distribution 1.

(**a**) $A(a)$ with $p = 0.5$.

(**b**) $A(a)$ with $p = 0.02$.
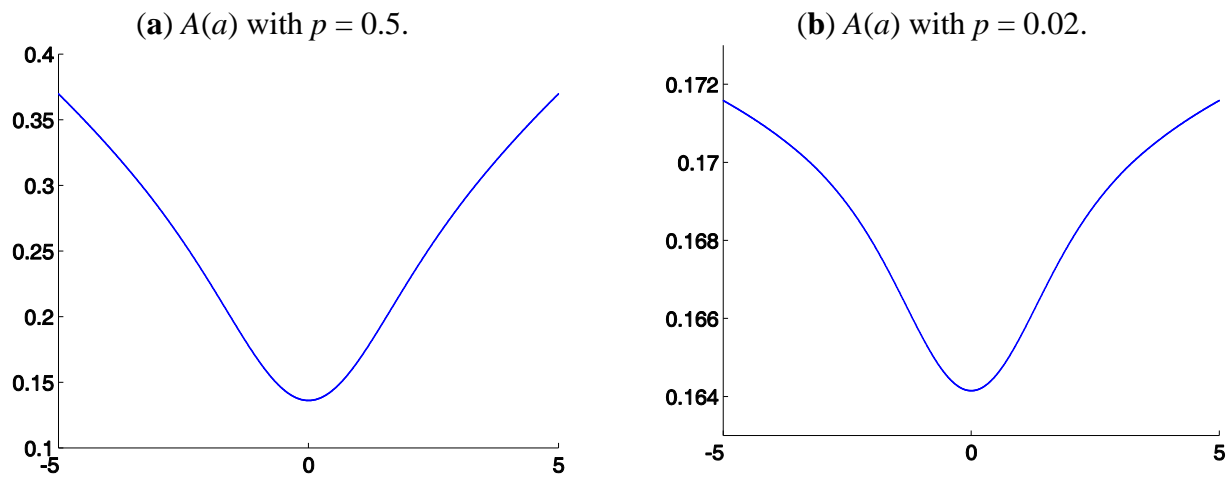
Figures 2. L^p averaging functional $A(a)$ for symmetric heavy-tailed Distribution 2 with $\alpha = 2$.

(**a**) $A(a)$ with $p = 0.5$.

(**b**) $A(a)$ with $p = 0.02$.

Figures 3. L^p averaging functional $A(a)$ for asymmetric heavy-tailed Distribution 3 with $\alpha = 2$.

(**a**) $A(a)$ with $p = 0.1$.

(**b**) $A(a)$ with $p = 0.02$.
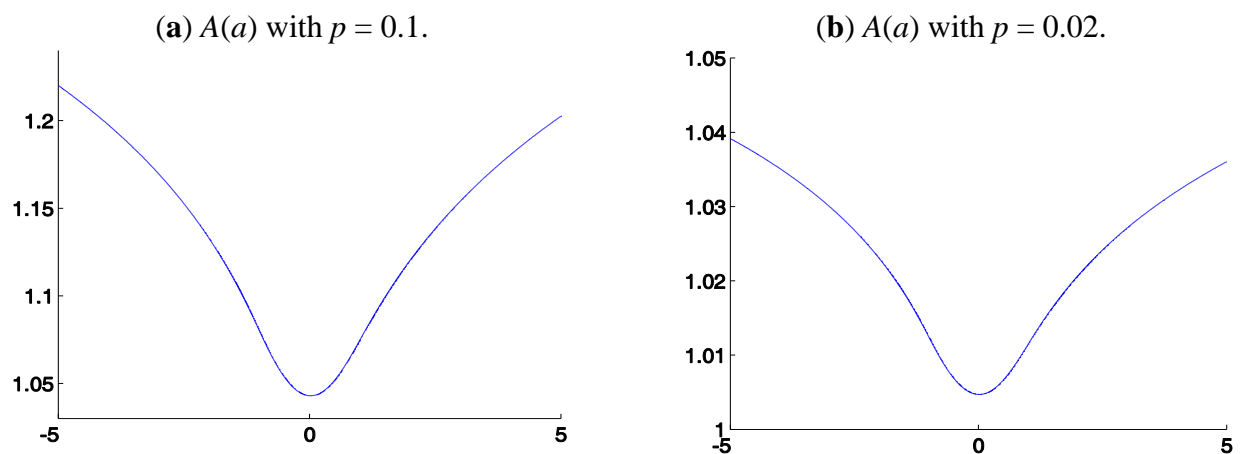
The structure of the L^p averaging functional $A(a)$ seen in Figures 1–3 and described in the previous paragraph occurs for all symmetric distributions, a situation that can be shown as follows. For symmetric distributions (that is, those for which $g(r) = f(r)$), the L^p averaging functional $A(a)$ can be written as

$$A(a) \;=\; \int_0^\infty (\, |r + a|^p + \, |r - a|^p \,) \, f(r) \, \mathrm{d}r \tag{9}$$

$A(a)$ is symmetric around $a = 0$, so we need consider only the behavior of $A(a)$ for $a \geq 0$. For $a \geq 0$,

$$\frac{\mathrm{d}A}{\mathrm{d}a}(a) \;=\; \int_0^\infty (\, (r + a)^p - \, |r - a|^p \,) \left(-\frac{\mathrm{d}f}{\mathrm{d}r}(r) \right) \mathrm{d}r \tag{10}$$

and

$$\frac{d^2 A}{\mathrm{d}a^2}(a) \;=\; \int_0^a p \, (\, (r + a)^{p-1} - (\, a - r \,)^{p-1}) \left(-\frac{\mathrm{d}f}{\mathrm{d}r}(r) \right) \mathrm{d}r$$

$$+ \int_a^\infty p \, (\, (r + a)^{p-1} + (\, r - a \,)^{p-1}) \left(-\frac{\mathrm{d}f}{\mathrm{d}r}(r) \right) \mathrm{d}r \tag{11}$$

One computes expressions (10) and (11) by differentiating the right sides of expressions (9) and (10), respectively, with respect to a. One expresses the integral to be differentiated as the sum of an integral on $(0,a)$ and an integral on (a,∞) and differentiates these two integrals separately. To simplify $\mathrm{d}A/\mathrm{d}a$ to the form given in (10), one integrates by parts and combines the two resulting integrals. From these expressions, one obtains first that $\mathrm{d}A/\mathrm{d}a(0) = 0$ and $\mathrm{d}^2A/\mathrm{d}a^2(0) > 0$, that is, there is a local minimum at $a = 0$ and second that, for all $a > 0$, $\mathrm{d}A/\mathrm{d}a(a) > 0$, that is, A is strictly monotonically increasing for $a > 0$. Thus, for symmetric pdfs, $A(a)$ has its global minimum at $a = 0$, that is, the L^p average exists and is equal to the mode of the distribution. There are no places where $\mathrm{d}A/\mathrm{d}a = 0$ other than at $a = 0$ and, on every finite closed interval that does not include the mode 0, $\mathrm{d}A/\mathrm{d}a$ is bounded away from 0. Standard line-search methods for calculating the minimum of this $A(a)$ are thus globally convergent.

A general analytical structure for asymmetric distributions analogous to that described above for symmetric distributions is not yet available because, for asymmetric distributions, the properties of $A(a)$ depend on additional properties of the probability density functions $f(r)$ and $g(r)$ that have not yet been clarified. Most of the previous statistical research about two-tailed distributions that extend infinitely in each direction has been focused on symmetric distributions and it is the symmetric case on which we will focus in the remainder of this paper.

3. l^p Averaging

It is meaningful to calculate an l^p average of a discrete set of data, that is, the point at which $B(a)$ achieves its global minimum, only for data from a distribution that satisfies Conditions (3) and for which the L^p average exists, that is, for which $0 < p < \beta - 1$. We propose the following algorithm.

Algorithm 1: Algorithm for l^p Averaging
   STEP 1. Sort the data x_i, $i = 1, 2, \ldots, I$, from smallest to largest. (To avoid proliferation of notation, use the same notation x_i, $i = 1, 2, \ldots, I$, for the data after sorting as before.)

STEP 2. Choose an integer q that represents the number of neighbors of a given point in the sorted data set in each direction (lower and higher index) that will be included in a local set of indices to be used in the "window" in Step 4. (The "window size" is thus $2q + 1$)

STEP 3. Choose a point x_j from which to start. (The median of the data, that is, the l^1 average, is generally a good choice for the initial x_j.)

STEP 4. For each $k, j - q \leq k \leq j + q$, calculate $B(x_k)$.

STEP 5. If the x_k that yields the minimum of the $B(x_k)$ calculated in Step 4 is x_j, stop. In this case, x_j is the computed l^p average of the data. Otherwise, let x_k be a new x_j and return to Step 4.

STEP 6. If convergence has not occurred within a predetermined number of iterations, stop and return an error message.

Remark 1. Algorithm 1 considers the values of $B(a)$ only at the data points x_i and not between data points. For a strictly between two consecutive data points x_i and x_{i+1}, $B(a)$ is concave and is above the line connecting $(x_i, B(x_i))$ and $(x_{i+1}, B(x_{i+1}))$, so a minimum cannot occur there. It is sufficient, therefore, to consider only the values of B at the points x_i when searching for a minimum. A graph of the points $(x_i, B(x_i))$, $i = 1, 2, \ldots, I$, approximates the graph of the continuum L^p functional $A(a)$, which, for symmetric distributions, has only one local minimum, namely, its global minimum. The graph of the points $(x_i, B(x_i))$ may have some relatively shallow local minima produced by the irregular spacing of the x_i (cf. Figures 4 below) and/or the asymmetry of the distribution. The window structure of Algorithm 1 is designed to allow the algorithm to "jump over" these local minima on its way to the global minimum.

Remark 2. The cost of Algorithm 1 is polynomial, namely, the cost $O(I \log I)$ of the sorting operation of Step 1 plus the cost of the iterations of Step 4, namely, $O(I^2)$ (= the number of iterations, which cannot exceed $O(I)$, times the cost $O(I)$ of calculating each iteration). Analogous algorithms for higher-dimensional averages are expected to retain this polynomial-time nature.

In computational experiments, we used samples of size $I = 2000$ from the symmetric heavy-tailed Distribution 2 with various α, $1 < \alpha \leq 3$, and window sizes $2q + 1 = 7, 9, 11, \ldots, 25$. For comparison with Figures 2, we present in Figures 4 the graphs of the points $(x_i, B(x_i))$ for the sample from Distribution 2 with $\alpha = 2$ and $p = 0.5$ and 0.02. The starting point for Step 3 of the Algorithm 1 was chosen to be x_{I-2q}, a point near the end of the right tail (beyond the limited domains shown in Figures 4). As mentioned in Step 3 of Algorithm 1, the median of the data is a much better choice for a starting point. However, choosing a point near the right tail makes the iterations of Algorithm 1 traverse a large distance before converging to an approximation of the l^p average and thus provides an excellent test for the robustness of Algorithm 1. Computational results for $p = 0.5, 0.1$ and 0.02 and for window sizes $2q + 1 = 7, 13, 19$ and 25 are presented in Tables 1–4. For reference, we note that the continuum L^p averages of Distribution 2, when they exist, that is, when $p < \alpha - 1$, are all 0. Thus, the errors of the l^p averages in Tables 1–4 are the same as the l^p averages themselves.

The entries in Tables 1–4 indicate that, for all cases with $p < \alpha - 1$, the l^p average computed by Algorithm 1 is an excellent approximant of the L^p average 0 given the large number of outliers and the huge spread of the data in Distribution 2. (For $\alpha = 3$ and $\alpha = 1.02$, the ranges of the data are $[-16.0, 22.6]$ and $[-6.44 \times 10^{154}, 5.02 \times 10^{169}]$, respectively. For $\alpha = 2, 1, 1.5, 1.1, 1.05, 1.04$ and 1.03, the ranges are between these two ranges.) The entries for $p = 0.5$ with $\alpha = 1.5$ and for $p = 0.1$ with

$\alpha = 1.1$, 1.05, 1.04 and 1.03 in Tables 1 and 2 indicate that, in a few cases when p is equal to or only slightly greater than $\alpha - 1$, the l^p average yielded by Algorithm 1 can still be a good approximant of the center of the distribution in spite of the fact that the l^p average is theoretically meaningful only when $p < \alpha - 1$. The entries for $p = 0.5$ with $\alpha = 1.1$, 1.05, 1.04, 1.03 and 1.02 and for $p = 0.1$ with $\alpha = 1.02$ indicate that, in accordance with expectations, when p is significantly greater than $\alpha - 1$, the l^p average produced by Algorithm 1 is not a meaningful approximant of the center of the distribution. Since larger window size is of assistance when attempting to "jump over" local minima, it is expected that l^p averages should converge to the L^p average 0 as the window size $2q + 1$ increases (and as the sample size increases). The results in Tables 1–4 confirm that, for the samples used in these calculations, increasing the window size does indeed increase the accuracy of the l^p averages as approximations of the L^p average 0. In addition, the results in Tables 3 and 4 for $p < \alpha - 1$ show that, for the samples used in these calculations, there is an optimal q, namely, $q = 19$ that produces l^p averages that are just as good as the l^p averages produced by the larger $q = 25$ but (due to smaller window size) requires less computational effort.

Figures 4. Points $(x_i, B(x_i))$ for 2000-point sample from symmetric heavy-tailed Distribution 2 with $\alpha = 2$.

(**a**) Points $(x_i, B(x_i))$ for $p = 0.5$.

(**b**) Points $(x_i, B(x_i))$ for $p = 0.02$.
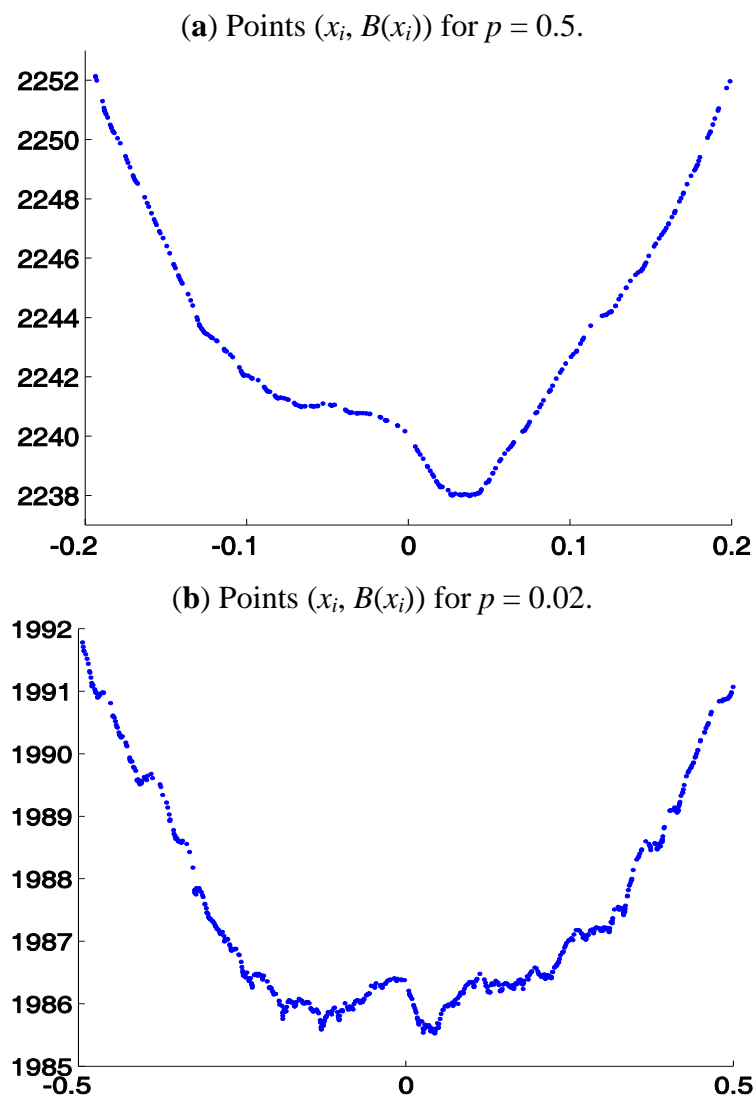
Algorithm 1 is applicable to heavy-tailed distributions in general but the rule for choosing q will certainly be dependent on the specific class of distributions under consideration. While this rule is not yet known precisely, we can provide here a description of the principles that will likely be the foundations for the rule. The choice of q is related to how wide the local minima in the discrete functional B are. The local minima of B occur at places where there are clusters of data points (due to expected statistical variation in the sample). Understanding the relationships between (1) the clustering properties of samples from the given class of distributions, (2) the widths of the local minima as functions of the clustering and (3) the p-dependent analytical properties of functional B will likely yield the rule for choosing q.

Table 1. Sample l^p averages calculated by Algorithm 1 with window size $2q + 1 = 7$ for 2000-point data set from Distribution 2.

α \ p	0.5	0.1	0.02
3	0.028	0.560	0.701
2	0.038	0.779	0.779
1.5	0.057	0.575	0.575
1.1	7.58	0.244	0.244
1.05	1.49×10^{30}	0.281	0.476
1.04	1.14×10^{45}	0.349	0.598
1.03	2.83×10^{74}	0.466	0.466
1.02	1.52×10^{119}	1.38×10^{16}	0.516

Table 2. Sample l^p averages calculated by Algorithm 1 with window size $2q + 1 = 13$ for 2000-point data set from Distribution 2.

α \ p	0.5	0.1	0.02
3	0.021	0.094	0.531
2	0.027	0.126	0.126
1.5	0.041	0.189	0.189
1.1	3.76	0.108	0.108
1.05	2.56×10^{29}	0.207	0.207
1.04	1.14×10^{45}	0.257	0.257
1.03	2.83×10^{74}	0.341	0.341
1.02	1.52×10^{119}	3.24×10^{14}	0.516

Table 3. Sample l^p averages calculated by Algorithm 1 with window size $2q + 1 = 19$ for 2000-point data set from Distribution 2.

α \ p	0.5	0.1	0.02
3	0.021	0.015	0.015
2	0.021	0.020	0.020
1.5	0.031	0.029	0.029
1.1	0.902	0.108	0.108
1.05	2.56×10^{29}	0.207	0.207
1.04	1.14×10^{45}	0.257	0.257
1.03	2.83×10^{74}	0.341	0.341
1.02	1.52×10^{119}	1.78×10^{7}	0.516

Table 4. Sample l^p averages calculated by Algorithm 1 with window size $2q + 1 = 25$ for 2000-point data set from Distribution 2.

α \ p	0.5	0.1	0.02
3	0.021	0.015	0.015
2	0.021	0.020	0.020
1.5	0.031	0.029	0.029
1.1	0.498	0.108	0.108
1.05	2.56×10^{29}	0. 207	0.207
1.04	1.14×10^{45}	0.257	0.257
1.03	2.83×10^{74}	0.341	0.341
1.02	1.52×10^{119}	2.37×10^6	0.516

4. Conclusions

The wide-spread impression that minimization of L^p and l^p functionals, $0 < p < 1$, is combinatorially expensive is valid for general situations in which no structure of the data is known. However, the results in this paper suggest that, when the data come from an appropriate statistical distribution, L^p and l^p averages can be calculated in polynomial time. The approach of the paper is applicable without precise knowledge of the parameters of the distribution. One does not need precise knowledge of the parameters but rather only generalizations of Conditions (3), an upper bound on the exponent $-\beta$ of the tail density and additional conditions for asymmetric distributions and for setting up a rule for choosing q in Algorithm 1.

Topics for future research include

- Quantitative rules for using information about the underlying continuum distribution to choose the q of Algorithm 1 based on a user's preferred tradeoff between maximum accuracy and minimum computational burden
- Investigation of the advantages and disadvantages of introducing smoothing in the $B(x_k)$ calculated in Step 4 of Algorithm 1 to increase the robustness against shallow local minima; connection of the smoothing with properties of the underlying distributions
- Description of the class(es) of symmetric and asymmetric univariate and multivariate distributions for which radially strictly monotonic L^p averaging functionals and radially nearly strictly monotonic l^p averaging functionals can be created and thus for which L^p and l^p averages can be calculated in polynomial time
- Investigation of convergence of the l^p average to the L^p average and of related issues of efficiency, optimality, breakdown point, influence function, *etc.*
- Investigation of the conditions under which L^p and l^p averages converge to the mode as $p \to 0$
- Treatment of more general univariate and multivariate l^p minimization problems including but not limited to l^p regression and matrix-constrained l^p minimization, for example, minimization of

$$\sum_{i=1}^I |x_i|^p \text{ subject to } A\mathbf{x} = \mathbf{b} \tag{12}$$

(cf. [17,18]) (The l^p averaging process considered in the present paper can be expressed in format (12).)

Many phenomena in human-based areas (sociology, cognitive science, psychology, economics, human networks, social media, *etc.*) are increasingly known to be represented by data that have large numbers of outliers and belong to very heavy-tailed distributions, which suggests that L^p and l^p averaging, L^p and l^p regression and more general L^p and l^p minimization tasks, $0 < p < 1$, will be important in practice. The results of the present paper provide the first indication that one may be able to solve, in polynomial time, generically combinatorially expensive L^p and l^p minimization problems for these phenomena by requiring only "natural" statistical structure without having to impose restrictions such as sparsity and without having to accept suboptimal local solutions instead of optimal global solutions.

Acknowledgment

References

1. Gribonval, R.; Nielsen, M. Sparse Approximations in Signal and Image Processing. *EURASIP Book Ser. Signal Process. Commun.* **2006**, *86*, 415–416.
2. Lai, M.-J.; Wang, J. An unconstrained ℓ_q minimization with $0 < q \leq 1$ for sparse solution of under-determined linear systems. *SIAM J. Optim.* **2010**, *21*, 82−101.
3. Chartrand, R. Exact reconstruction of sparse signals via nonconvex minimization. *IEEE Signal Process. Lett.* **2007**, *14*, 707–710.
4. Candès, E.J.; Wakin, M.B. An introduction to compressive sampling. *IEEE Signal Process. Mag.* **2008**, *25*, 21–30.
5. Auquiert, P.; Gibaru, O.; Nyiri, E. Fast L_1-C^k polynomial spline interpolation algorithm with shape-preserving properties. *Comput. Aided Geom. Design* **2011**, *28*, 65–74.
6. Yu, L.; Jin, Q.; Lavery, J.E.; Fang, S.-C. Univariate cubic L_1 interpolating splines: Spline functional, window size and analysis-based algorithm. *Algorithms* **2010**, *3*, 311–328.
7. Candès, E.J.; Li, X.; Ma, Y.; Wright, J. Robust principal component analysis? *J. ACM* **2011**, *58*, 1–37.
8. Ke, Q.; Kanade, T. Robust L_1 norm factorization in the presence of outliers and missing data by alternative convex programming. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, San Diego, CA, USA, 20−25 June 2005; Schmid, C., Soatto, S., Tomasi, C., Eds.; IEEE Computer Society: Los Alamitos, CA, USA, 2005; pp. 739–746.
9. Kwak, N. Principal component analysis based on $L1$-norm maximization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 1672–1680.
10. Dodge, Y. Statistical Data Analysis Based on the L_1 Norm and Related Methods. In *Proceedings of the Conference on Statistical Data Analysis Based on the L_1 Norm and Related Methods*, Neuchâtel, Switzerland, 4–9 August 2002; Birkhäuser: Basel, Switzerland, 2002.

11. Nolan, J.P. Multivariate stable distributions: Approximation, estimation, simulation and identification. In *A Practical Guide to Heavy Tails*; Adler, R.J., Feldman, R.E., Taqqu, M.S., Eds.; Birkhäuser: Cambridge, MA, USA, 1998; pp. 509–525.

12. Resnick, S.I. *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*; Springer-Verlag: Berlin, Germany, 2007.

13. Faloutsos, M.; Faloutsos, P.; Faloutsos, C. On power-law relationships of the internet topology. *Comp. Comm. Rev.* **1999**, *29*, 251–262.

14. Willinger, W.; Govindan, R.; Jamin, S.; Paxson, V.; Shenker, S. Scaling phenomena in the Internet: Critically examining criticality. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 2573–2580.

15. Rachev, S.T.; Menn, C.; Fabozzi, F.J. *Fat-Tailed and Skewed Asset Return Distributions: Implications for Risk Management, Portfolio Selection, and Option Pricing*; John Wiley: Hoboken, NJ, USA, 2005.

16. Reed, W.J.; Jorgensen, M.A. The double Pareto-lognormal distribution—A new parametric model for size distributions. *Comm. Statist. Theory Methods* **2004**, *33*, 1733–1753.

17. Foucart, S.; Lai, M.-J. Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$. *Appl. Comput. Harmon. Anal.* **2009**, *26*, 395–407.

18. Ge, D.; Jiang, X.; Ye, Y. A note on the complexity of L_p minimization. *Math. Program.* **2011**, *129*, 285–299.

19. Gribonval, R.; Nielsen, M. Highly sparse representations from dictionaries are unique and independent of the sparseness measure. *Appl. Comput. Harmon. Anal.* **2007**, *22*, 335–355.

20. Wang, M.; Xu, W.; Tang, A. On the Performance of Sparse Recovery via L_p-minimization ($0 \leq p \leq 1$), *IEEE Trans. Info. Theory* **2011**, *57*, 7255–7278.

21. Wilcox, R.R. *Introduction to Robust Estimation and Hypothesis Testing*, 2nd ed.; Elsevier: Burlington, MA, USA, 2005.